IJDC | Conference Pre-print

Data Sources and Persistent Identifiers in the Open Science Research Graph of OpenAIRE

Jochen Schirrwagen
Bielefeld University, Germany

Alessia Bardi
ISTI, CNR, Pisa, Italy

Andreas Czerniak
Bielefeld University, Germany

Aenne Loehden
Bielefeld University, Germany

Najla Rettberg
University of Göttingen, Germany

Mike Mertens
University of Göttingen, Germany

Paolo Manghi
ISTI, CNR, Pisa, Italy

Abstract

In this article, we give an overview of the data source typologies used in OpenAIRE and provide an outline on the role of persistent identifiers in the aggregation, curation and provision workflows that lead to the generation of the Research Graph in OpenAIRE.

Submitted 16 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to. Jochen Schirrwagen, University Library Universität Bielefeld, Bielefelder IT-Servicezentrum, Postfach 10 01 31, D-33501 Bielefeld, Germany. Email: jochen.schirrwagen@uni-bielefeld.de

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

OpenAIRE (www.openaire.eu) is the European infrastructure for Open Science in Europe. Its mission is closely linked to the mission of the European Commission: to provide unlimited, barrier free, open access to research outputs financed by public funding in Europe.
OpenAIRE supports the implementation and alignment of Open Science policies at the international level by developing and promoting the adoption of global open standards and interoperability guidelines (guidelines.openaire.eu) to realize a sustainable, participatory, trusted, scholarly communication ecosystem, open to all relevant stakeholders (e.g. research communities, funders, project coordinators) and capable of engaging society and foster innovation.
Thanks to the network of 34 National Open Access Desks (NOADs), OpenAIRE supports the implementation of Open Science at the local and national level, supporting researchers, project coordinators, funders and policy makers with training and support activities like workshops and webinars.
With its technical infrastructure, OpenAIRE materializes an open, de-duplicated, participatory metadata research graph of interlinked objects of the research life-cycle (including research literature, datasets, software and projects). The graph is materialized by collecting metadata records from thousands of sources of different types. On OpenAIRE's beta infrastructure (beta.explore.openaire.eu) we provide a preview of the graph which is composed of more than 100 millions of metadata records collected from more than 9,000 scholarly data sources world-wide.
This paper will outline all the types of data sources that OpenAIRE gathers from and will focus on the importance of persistent identifiers as enabling technology for metadata curation, enrichment, monitoring and value-added services.

Data Source Typologies in OpenAIRE

OpenAIRE is unique in terms of integrating heterogeneous data sources that provide directory services, services exposing metadata records about scholarly objects and research information, and services that provide relationships among them. We distinguish between the following types of data sources:

- Institutional or thematic repositories: Information systems where scientists upload the bibliographic metadata and deposit full-texts of their articles, due to obligations from their organization or due to community practices (e.g. ArXiv, Europe PMC);
- Open Access Publishers and journals: Information system of open access publishers or relative journals, which offer bibliographic metadata and full-texts of their published articles;
- Data repositories/archives: Information systems where scientists deposit descriptive metadata and files about their research data (also known as scientific data, datasets, etc.).;
- Software repositories: Information systems where scientists deposit descriptive metadata and files about their research software and tools;
- Hybrid repositories/archives: information systems where scientists deposit metadata and files of any kind of scientific products, including scientific literature, research data and research software (e.g. Zenodo)

- Aggregator services: Information systems that collect descriptive metadata records from multiple sources. Examples are BASE, DOAJ;
- Current Research Information Systems (CRIS): Information systems adopted by research and academic organizations to keep track of their research administration records and relative results; examples of CRIS content are articles or datasets funded by projects, their principal investigators, facilities acquired thanks to funding, etc..
- Funder databases: databases managed by funders from which it is possible to obtain an authoritative list of funded research projects
- Directories of data sources (DoDS): information systems created with the intent of maintaining authoritative registries on scholarly communication sources, such as OpenDOAR for open access repositories, re3data.org for data repositories, DOAJ for open access journals and DRIS for research information systems.
- Other types of sources, which provide metadata descriptions of different types of scholarly objects like ORCID for researchers and relative outputs, ScholeXplorer for links between research data and scientific literature, and CrossRef for scientific literature and relative DOIs.

Table 1 summarizes the number of sources aggregated by OpenAIRE by July 2019, grouped by typology, on the production infrastructure.

Table 1. Number of aggregated data sources and their types (July-2019).

Datasource type	Quantity
Journals	15442
Institutional Repositories	970
Data Repositories	159
Publication Repository Aggregators	63
Thematic Repositories	62
Funder Databases	21
Publication Catalogue	13
Software Repositories	7
CRIS	5
Data Repository Aggregators	3
Data Source Registries	3
PID Registries	3

Persistent Identifiers for Metadata Management and Value-Added Services

The OpenAIRE research graph data model [CITATION Man19 \l 1031] relies on the persistent identification of all its entities it is made of.
The integration workflow of data sources and their content in order to build the research graph can be divided into the following steps as depicted in Figure 1.

Figure 1. Aggregation and content provision steps

OpenAIRE interfaces with Directories of Data Sources whose metadata and data source identifiers are used for the **registration** of data sources in the OpenAIRE infrastructure. Relying on such identifiers, e.g. service-DOIs for data repositories, ISSN for journals or OpenDOAR-IDs for institutional repositories, avoids redundant registrations and serves as provenance information in metadata records.
OpenAIRE has issued Guidelines for certain types of data sources which are aimed to provide orientation to repository managers to describe bibliographic records using established metadata standards and controlled vocabularies, such as Dublin Core, DataCite and CERIF-XML. In particular the Guidelines recommend the use of persistent identifiers not only for the scientific product itself (e.g. DOI, URN, PMID) and the medium it was published in (e.g. ISSN of a journal, ISBN of a book), but also for authors and contributors (e.g. ORCID), funding organisations (e.g. fundRef) and projects.
In the aggregation process of bibliographic metadata records, information on PIDs are extracted and harmonized so that they can be used for subsequent curation, enrichment and contextualization and finally the publishing tasks.
The roles of PIDs in OpenAIRE is not limited to the identification of data sources and scholarly objects. As depicted in Figure 2 PIDs are used in the enrichment stage in two ways.
By enriching via PIDs:

- adding records of cited publications not contained in the metadata graph, by automatically gathering metadata exposed by resolvers;
- adding records of publications not contained in the metadata information space so far, by manual inclusion by users.

By enriching of PIDs:

- adding PIDs not contained or not explicitly contained in the metadata information space so far, within aggregation/curation;
- enriching of PIDs: adding PIDs not contained or not explicitly contained in a data source's metadata so far.

The availability of PIDs is crucial also for the deduplication process that OpenAIRE applies to identify and merge duplicates of publications, datasets and software [CITATION Atz18 \l 1031]. In particular, if the PIDs of the two records are the same, then the two records are duplicates. Otherwise, the algorithm continues checking other available metadata fields. In particular, for publications, the algorithm checks:

- If the titles of the two records contain numbers and these numbers are not the same, then the records are not duplicates (e.g. "Annual Report 2019" is not the same as "Annual Report 2018");
- If the two records contain different numbers of authors, then the records are not duplicates.
- If the two records have not yet identified as different by the previous checks, then a final condition on titles is applied: titles are normalised and compared for similarity by applying the Levenstein distance algorithm. The algorithm returns a number in the range [0,1], where 0 means "very different" and 1 means "equal". If the distance is greater than or equal 0,99 the two records are identified as duplicates.

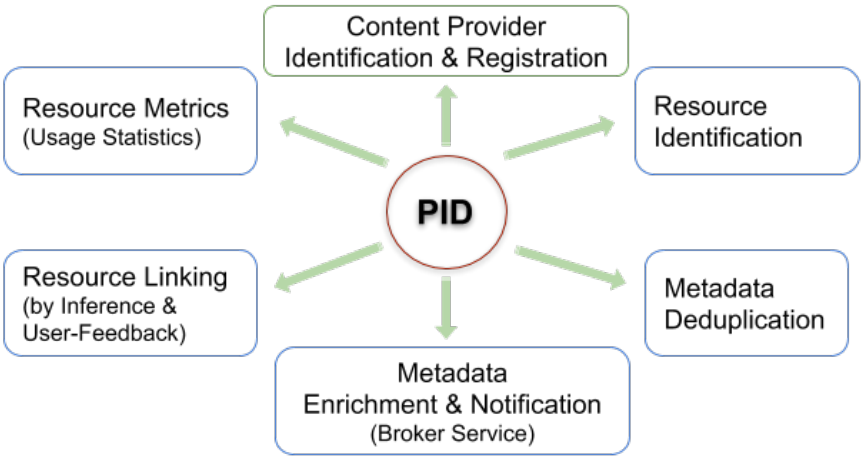

Figure 2. The different roles of PIDs in OpenAIRE service functionalities

Acknowledgements

This research was funded by the EC OpenAIRE-Advance project (grant 777541), call H2020-EINFRA-2017-1.

References

Atzori, C., Manghi, P., & Bardi, A. (2018). GDup: De-Duplication of Scholarly Communication Big Graphs. *IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)* (pp. 142-151). IEEE. doi: 10.1109/BDCAT.2018.00025

Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., & Principe, P. (2019). *The OpenAIRE Research Graph Data Model.* doi: 10.5281/zenodo.2643199